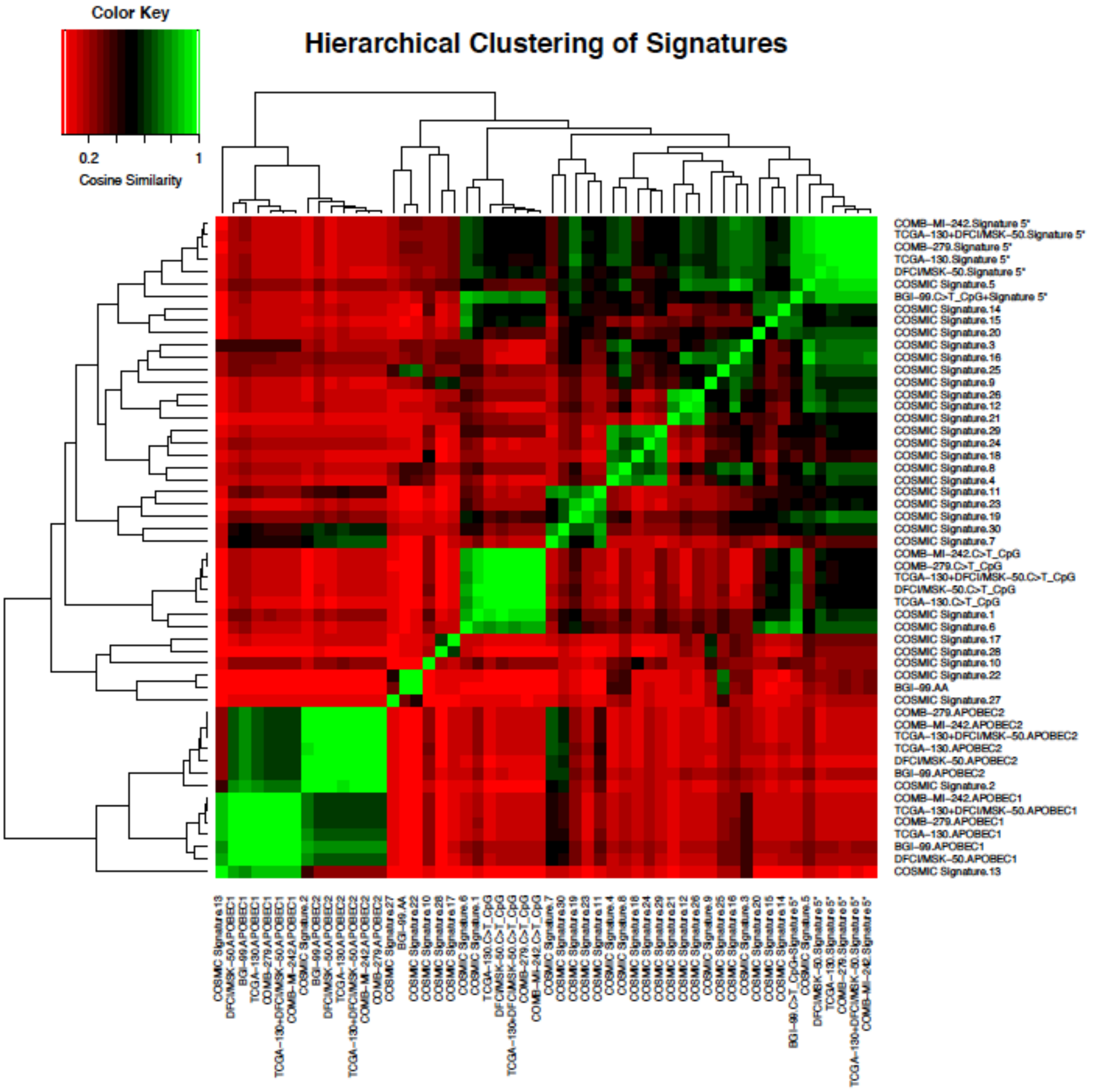
Somatic *ERCC2* Mutations Are Associated with a Distinct Genomic Signature in Urothelial Tumors

Jaegil Kim, Kent W Mouw, Paz Polak, Lior Z Braunstein, Atanas Kamburov, David J Kwiatkowski, Jonathan E Rosenberg, Eliezer M Van Allen, Alan D'Andrea, Gad Getz

SUPPLEMENTARY INFORMATION

Contents

Supplementary Figure 1a



Signature 5*

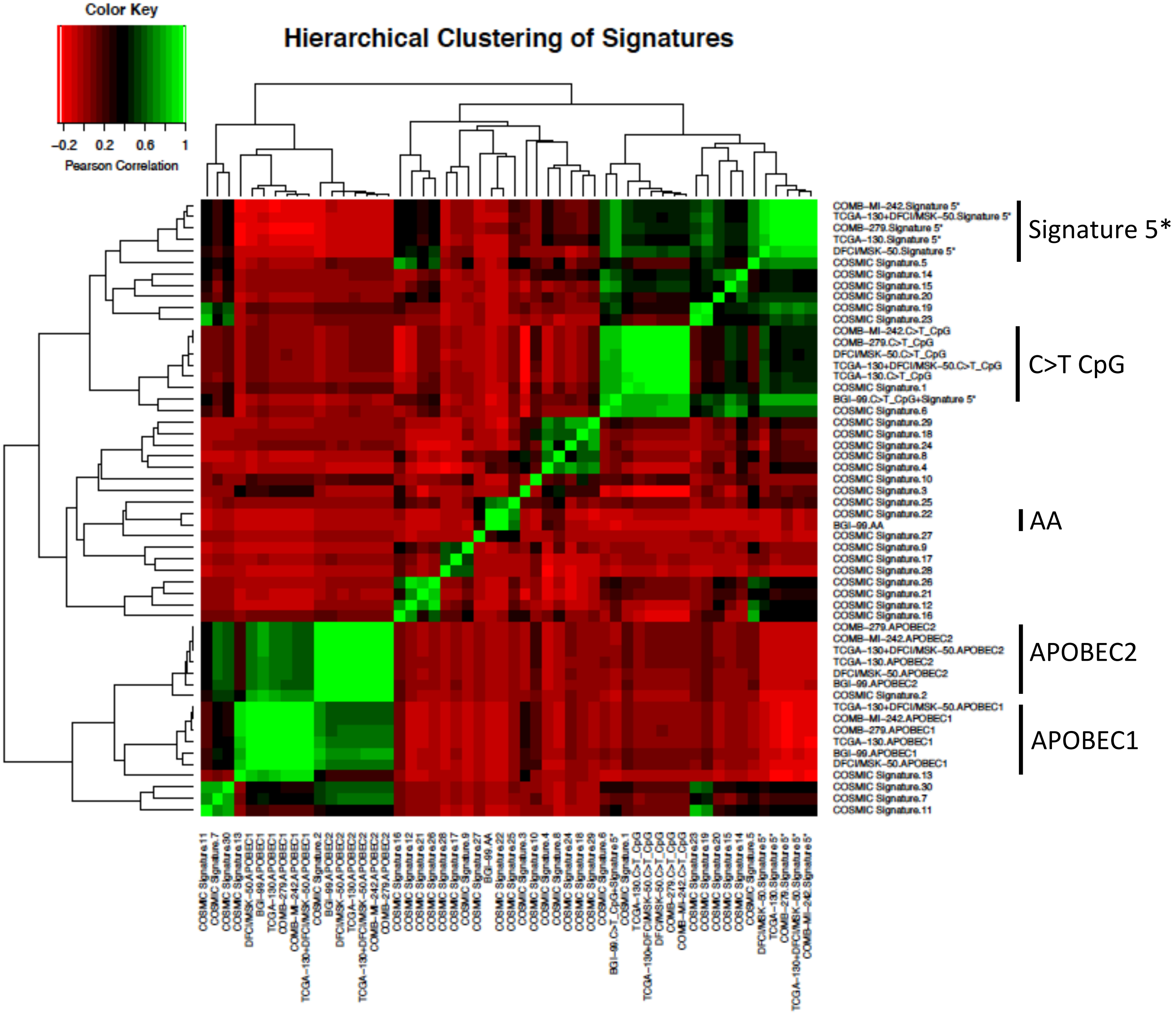
C>T CpG

AA

APOBEC2

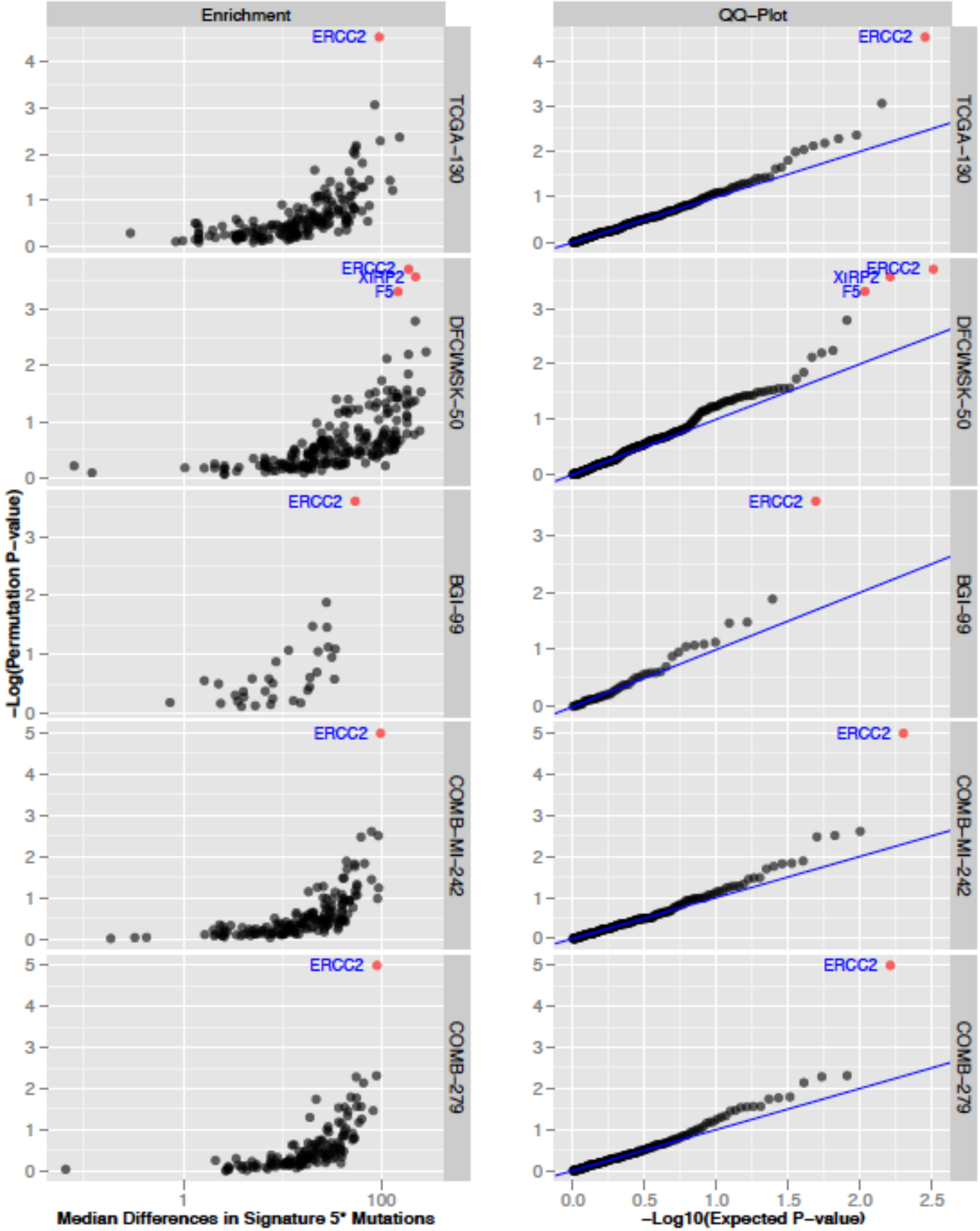
APOBEC1

Supplementary Figure 1b



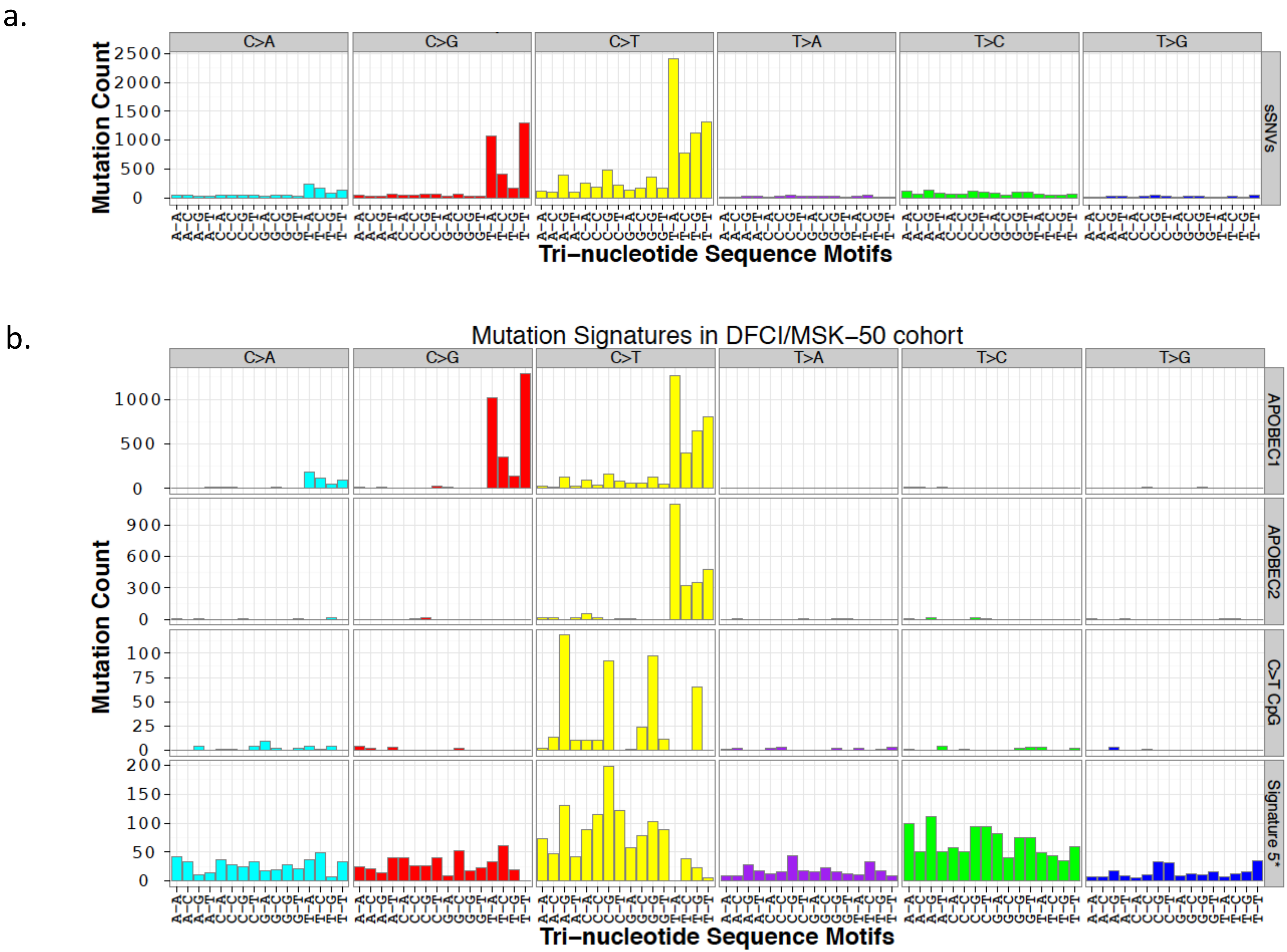
Supplementary Figure 1 Unsupervised hierarchical clustering of signatures identified in three urothelial tumor cohorts (TCGA-130, DFCI/MSK-50, and BGI-99) and the 30 signatures described by COSMIC.[1, 2] COMB-279 is the combined cohort including all tumors from the TCGA-130, DFCI/MSK-50, and BGI-99 cohorts. COMB-MI-242 are all muscle-invasive tumors from the three cohorts. AA: Aristolochic acid. **(a)** Cosine similarity. **(b)** Pearson correlation.

Supplementary Figure 2



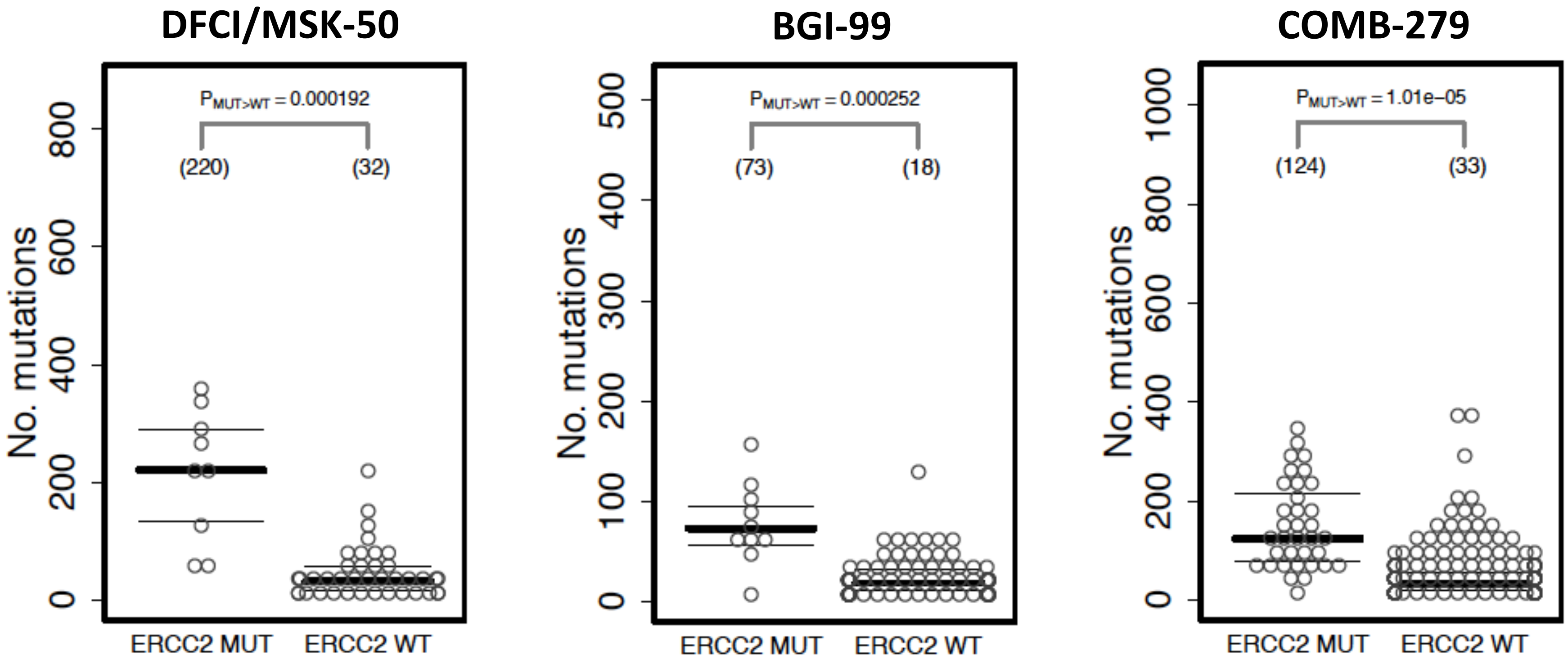
Supplementary Figure 2 Summary of mutation enrichment analyses for signature 5* across cohorts. All genes mutated in >5% of tumors in the cohort were included in the analysis. A permutation-based method was applied to account for the overall number of non-silent mutations per sample and per gene (**Methods**).[3] Q-Q plots show observed versus expected p-values for each of the analyses. Genes with Benjamini-Hochberg False Discovery Rate Q<0.1 are shown in red and labeled. *ERCC2* was the only gene that was significant in each of the cohorts. COMB-279: all 279 tumors across the 3 cohorts. COMB-MI-242: all 242 muscle-invasive tumors across the 3 cohorts.

Supplementary Figure 3



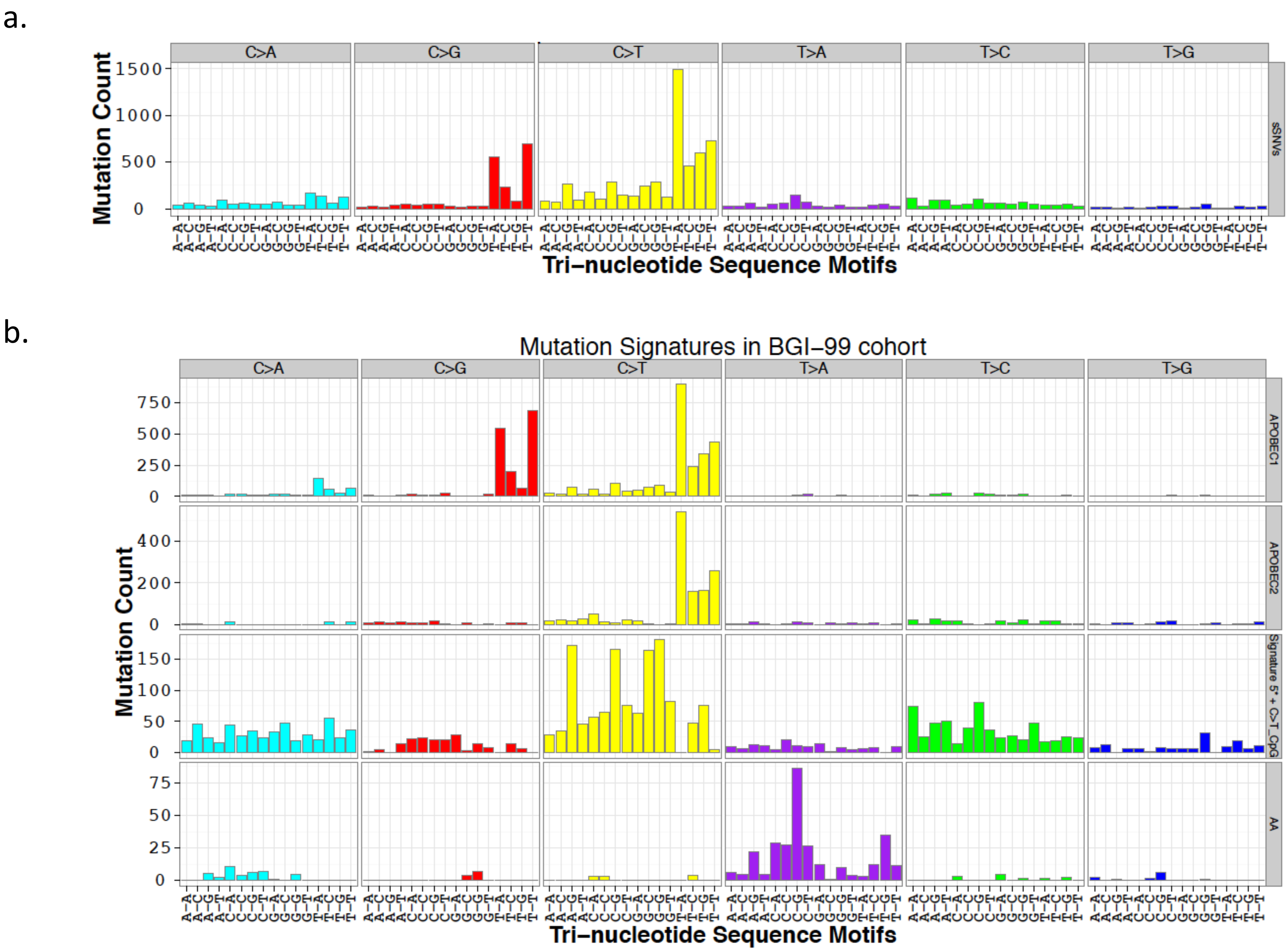
Supplementary Figure 3 Mutational signature analysis of the DFCI/MSK-50 cohort. (**a**) The spectrum of base changes identified in the DFCI/MSK-50 cohort displayed as the mutated pyrimidine and the adjacent 3' and 5' bases. sSNV: somatic single nucleotide variations. (**b**) A Bayesian non-negative matrix factorization algorithm was applied to identify signatures from the overall mutation spectrum. Four distinct mutational processes were identified that closely resemble the signatures identified in the TCGA-130 cohort in Figure 1b.

Supplementary Figure 4



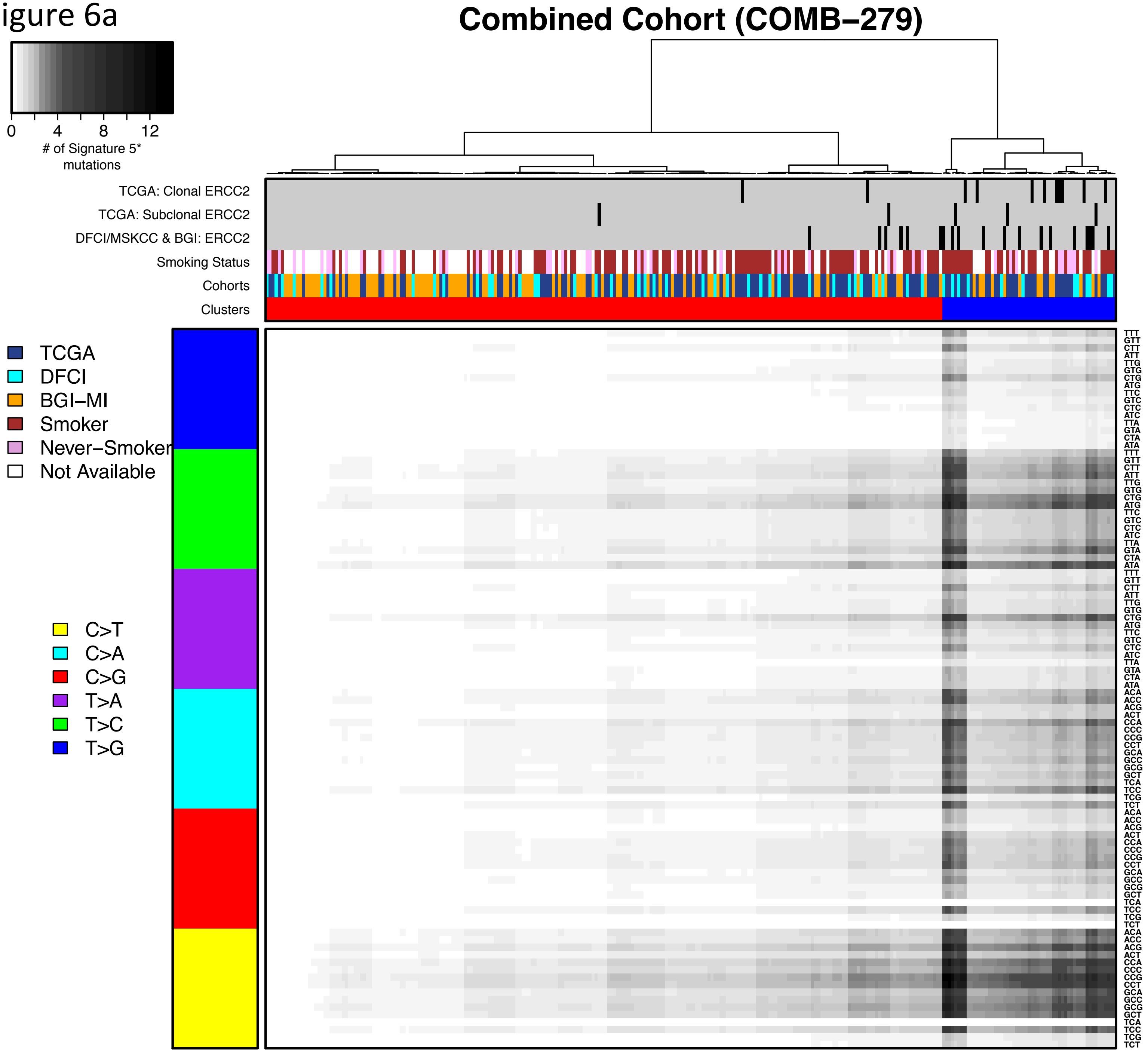
Supplementary Figure 4 Comparison of signature 5* activity in tumors with mutant versus WT *ERCC2* in the DFCI/MSK-50, BGI-99, and combined (COMB-279 = TCGA-130 + DFCI/MSK-50 + BGI-99) cohorts. The median estimated number of mutations is shown in parentheses. One-sided p-values were calculated using a permutation-based method that maintains the overall number of non-silent mutations per sample and per gene (**Methods**).[3] Note that the BGI-99 signature includes a contribution from the C>T CpG signature.

Supplementary Figure 5

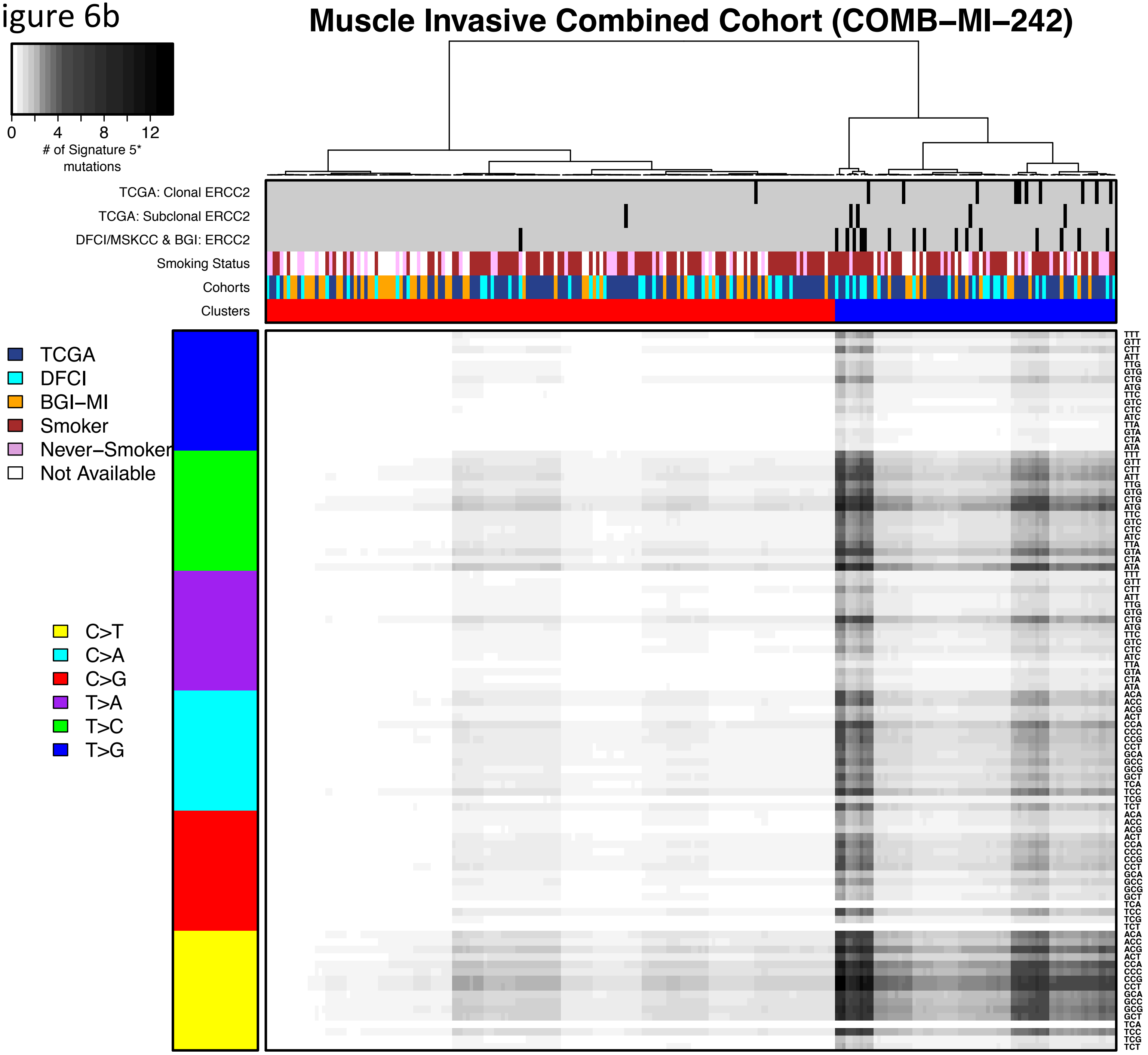


Supplementary Figure 5 Mutational signature analysis of the BGI-99 cohort. (**a**) The spectrum of base changes identified in the BGI-99 cohort displayed as the mutated pyrimidine and the adjacent 3’ and 5’ bases. (**b**) A Bayesian non-negative matrix factorization algorithm was applied to identify signatures from the overall mutation spectrum. Four distinct mutational signatures were identified: two signatures resembling those attributed to APOBEC activity (also seen in the TCGA-130 and DFCI/MSK-50 cohorts), a signature attributed to aristolochic acid (AA) exposure, and a signature representing the superposition of the C>T CpG signature and signature 5*.

Supplementary Figure 6a

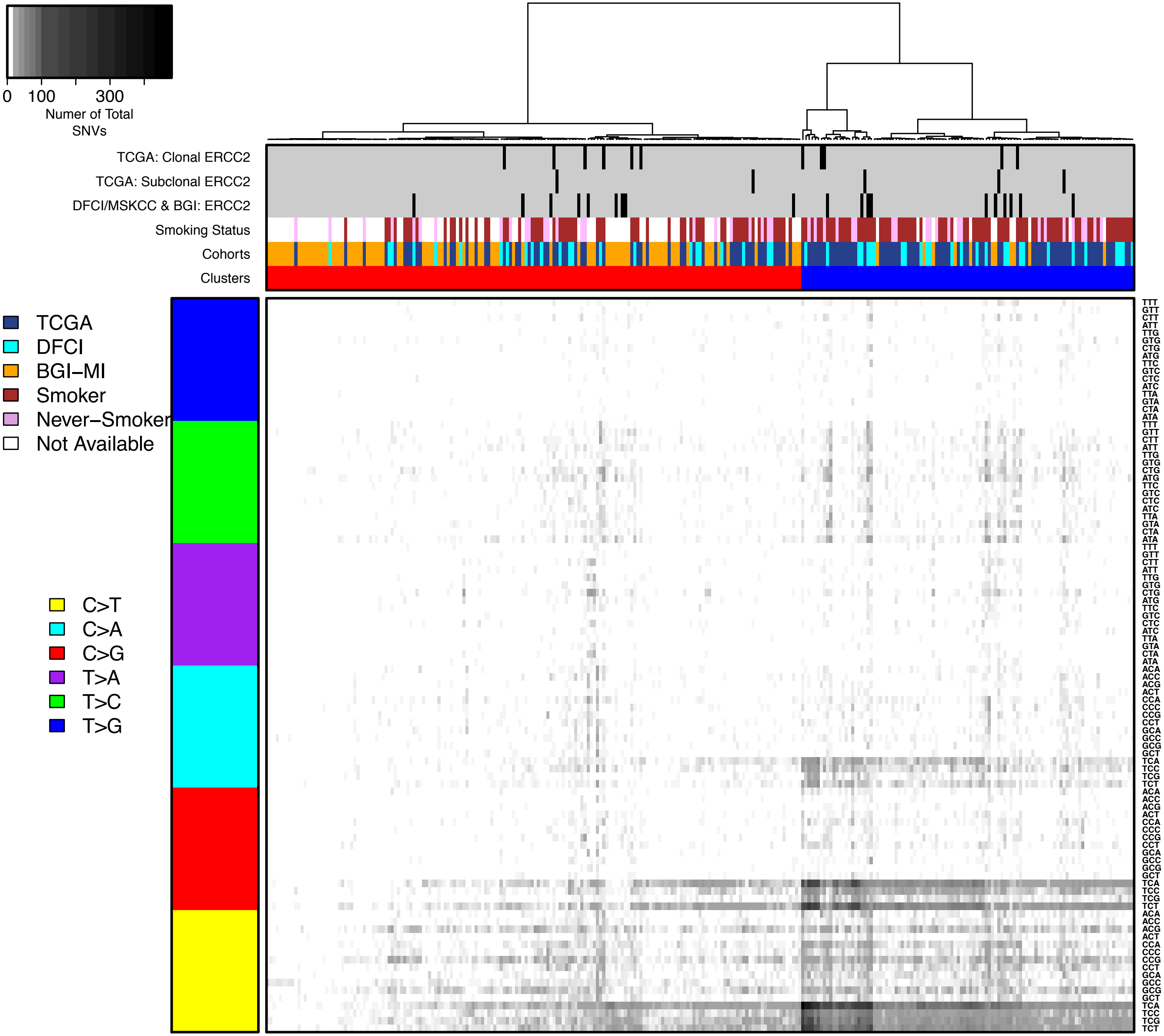


Supplementary Figure 6b

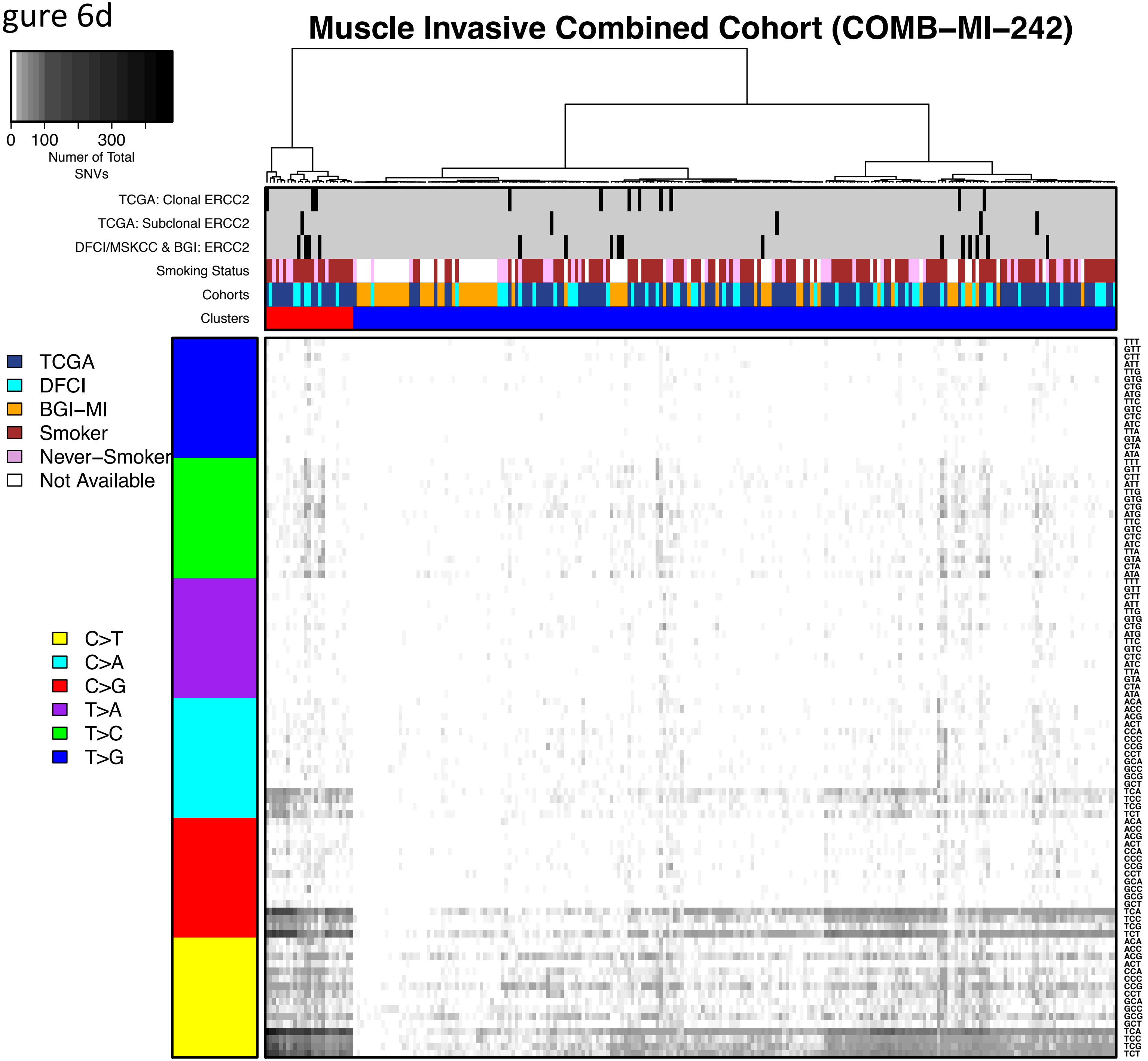


Supplementary Figure 6c

Combined Cohort (COMB–279)

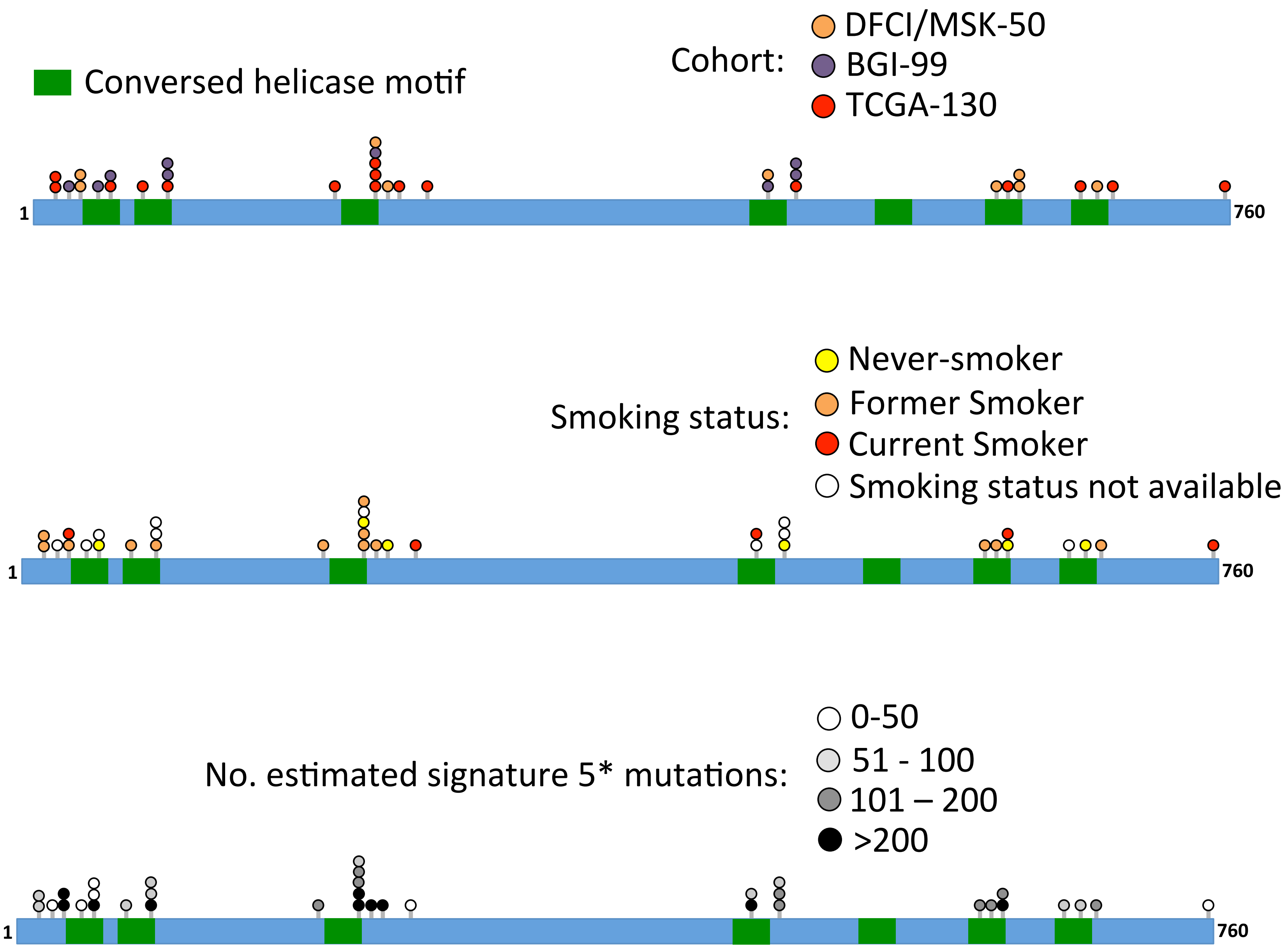


Supplementary Figure 6d

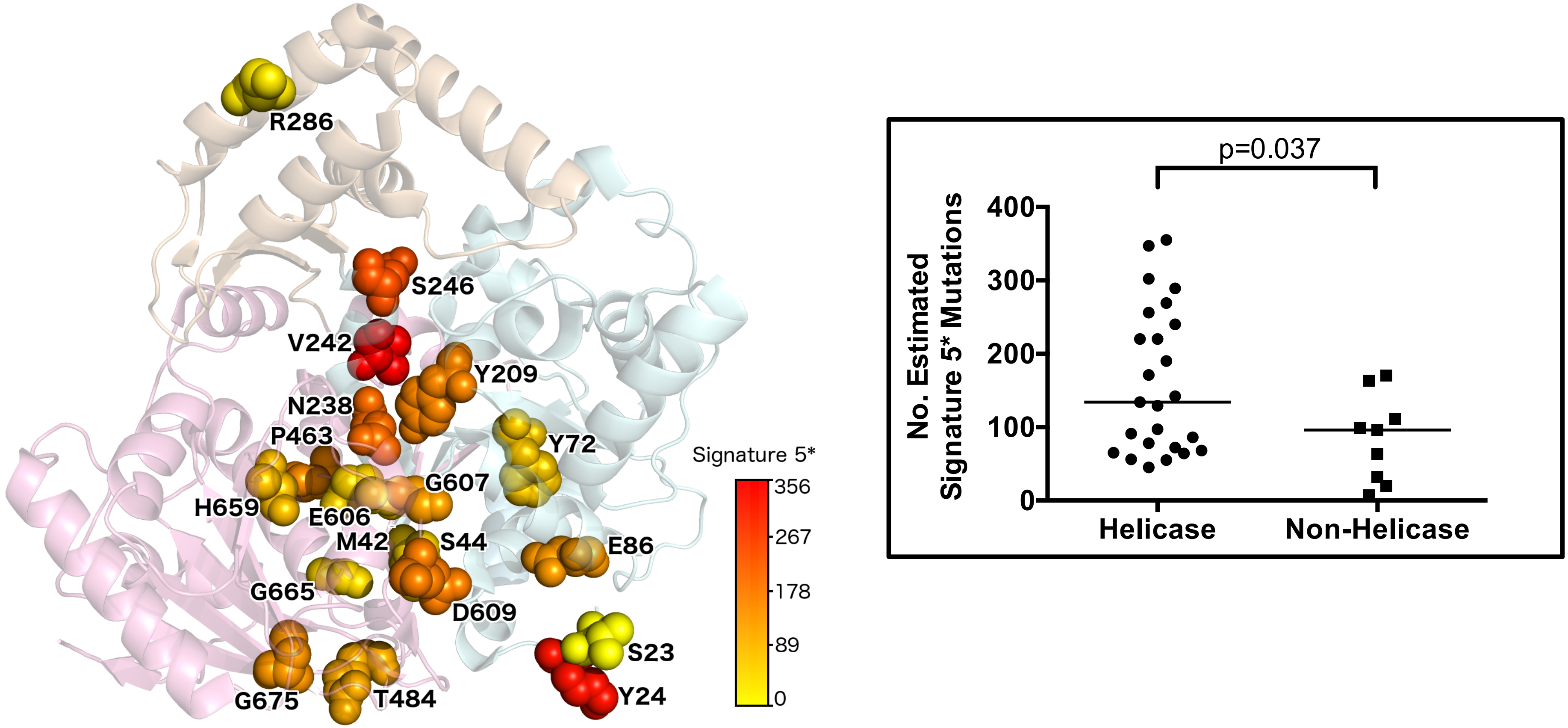


Supplementary Figure 6 Unsupervised hierarchical clustering analyses. **(a)** Clustering of signature 5* activity (as 96 trinucleotide mutational contexts) in the combined (COMB-279) cohort. Separate tracks are included for clonal (defined as probability[cancer cell fraction≥0.95] >0.5) and subclonal *ERCC2* mutations for the TCGA-130 cohort. Tumors segregated into two clusters of 222 (shown in red) and 57 (shown in blue) tumors. Twenty-five of the 35 *ERCC2* mutated tumors belonged to the second (blue) cluster (P=1.7x10^{-12}, two-tailed Fisher's exact test). (**b)** Clustering of signature 5* activity was also performed for all 242 muscle-invasive tumors from the combined cohort (COMB-MI-242). Tumors segregated into two clusters of 162 (red) and 80 (blue) tumors. Twenty-nine of 32 *ERCC2* mutated tumors belonged to the second (blue) cluster (P=4.4x10^{-14}). **(c)** Clustering of all non-silent SNVs in the COMB-279 cohort segregated tumors into clusters of 172 (red) and 107 (blue) tumors. Eighteen of 35 *ERCC2* mutated tumors belonged to the second (blue) cluster (P=0.1). **(d)** Clustering of all non-silent SNVs in the COMB-MI-242 cohort segregated tumors into clusters of 25 (red) and 217 (blue) tumors. Twenty-four of 32 *ERCC2* mutated tumors belonged to the second (blue) cluster (P=0.008).

Supplementary Figure 7a

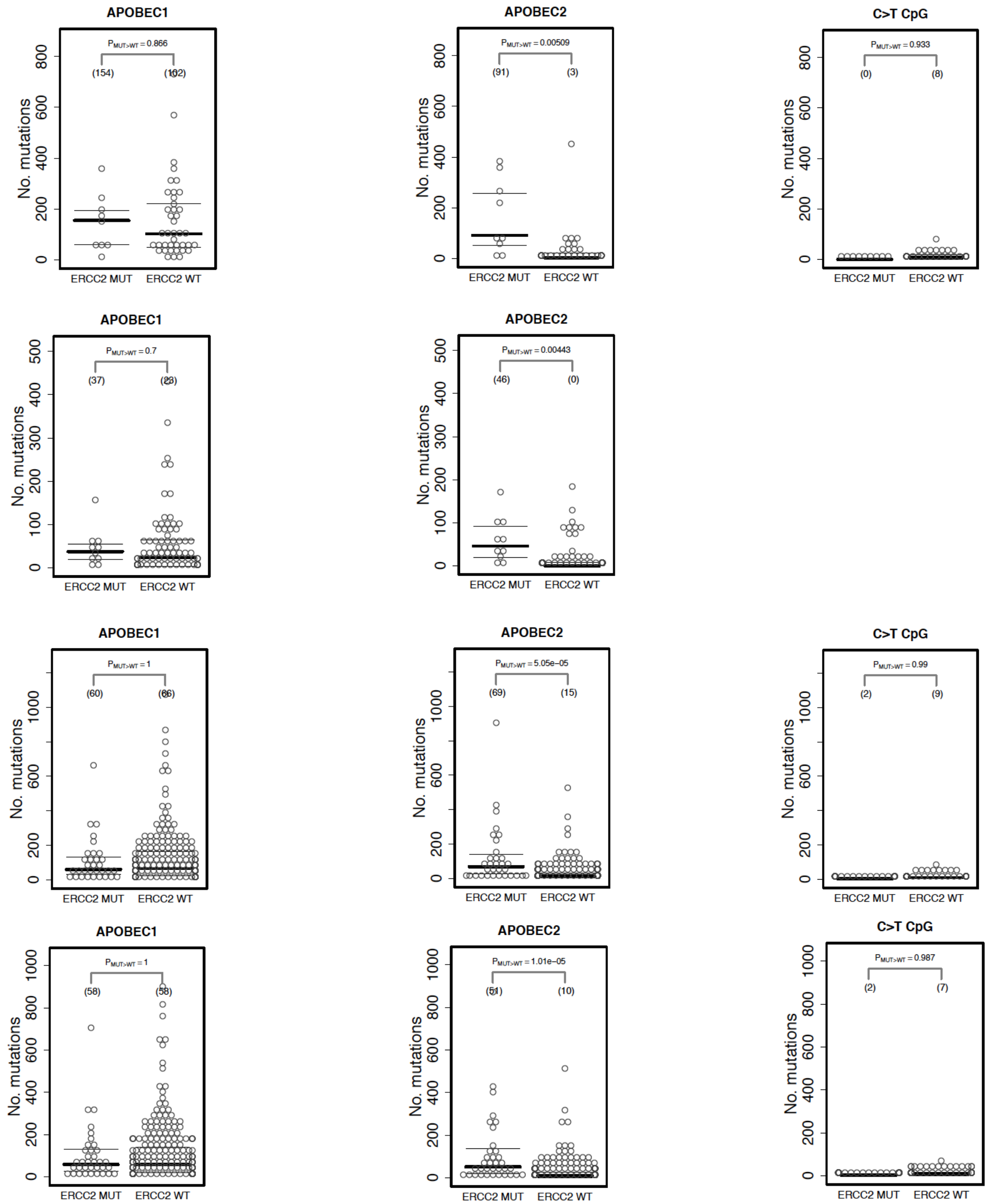


Supplementary Figure 7b



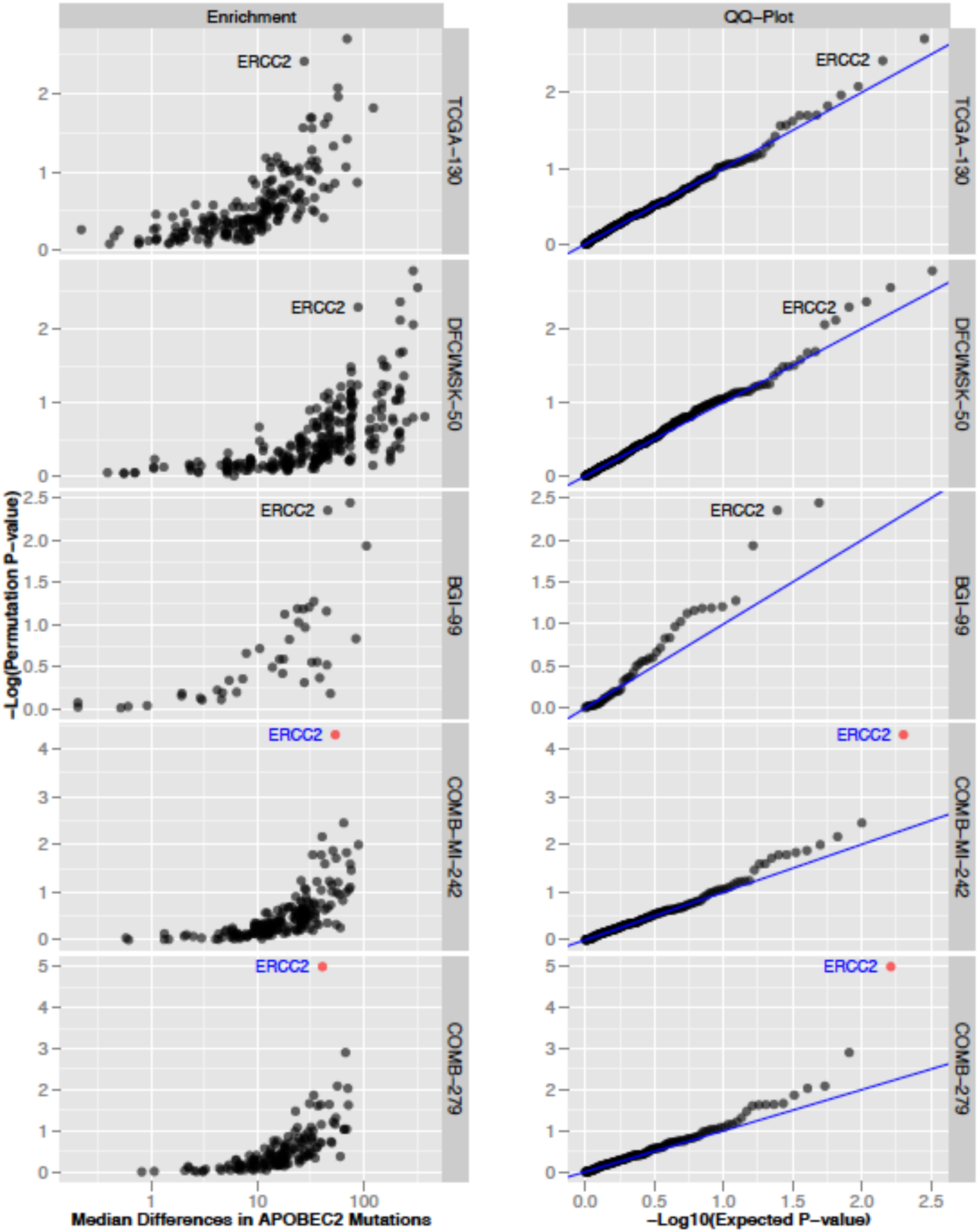
Supplementary Figure 7 Location and properties of *ERCC2* missense mutations. **(a)** The location of all missense mutations observed across cohorts are plotted along the length of the *ERCC2* gene (760 amino acids) by cohort, smoking status, and signature 5* activity. Mutations cluster within, or adjacent to, conserved helicase motifs (shown in green). One splice site mutation was present in a tumor from the BGI-99 cohort and is not shown here; all other mutations were missense mutations. **(b)** *ERCC2* missense mutations mapped to their predicted equivalent location on an archaeabacterial *ERCC2* crystal structure (PDB ID: 3CRV) and color-coded by estimated number of signature 5* mutations (**Methods**). For amino acids mutated in more than one tumor, the average number of signature 5* mutations is shown. Helicase domains are shaded in pink and green. Mutations located within or adjacent to (≤10 amino acids) the conserved helicase motifs were associated with a significantly higher number of signature 5* mutations compared to mutations located elsewhere in the protein (P=0.037). CLUMPS analysis revealed significant spatial clustering of mutations within the 3D structure (P=0.0026).[4]

Supplementary Figure 8



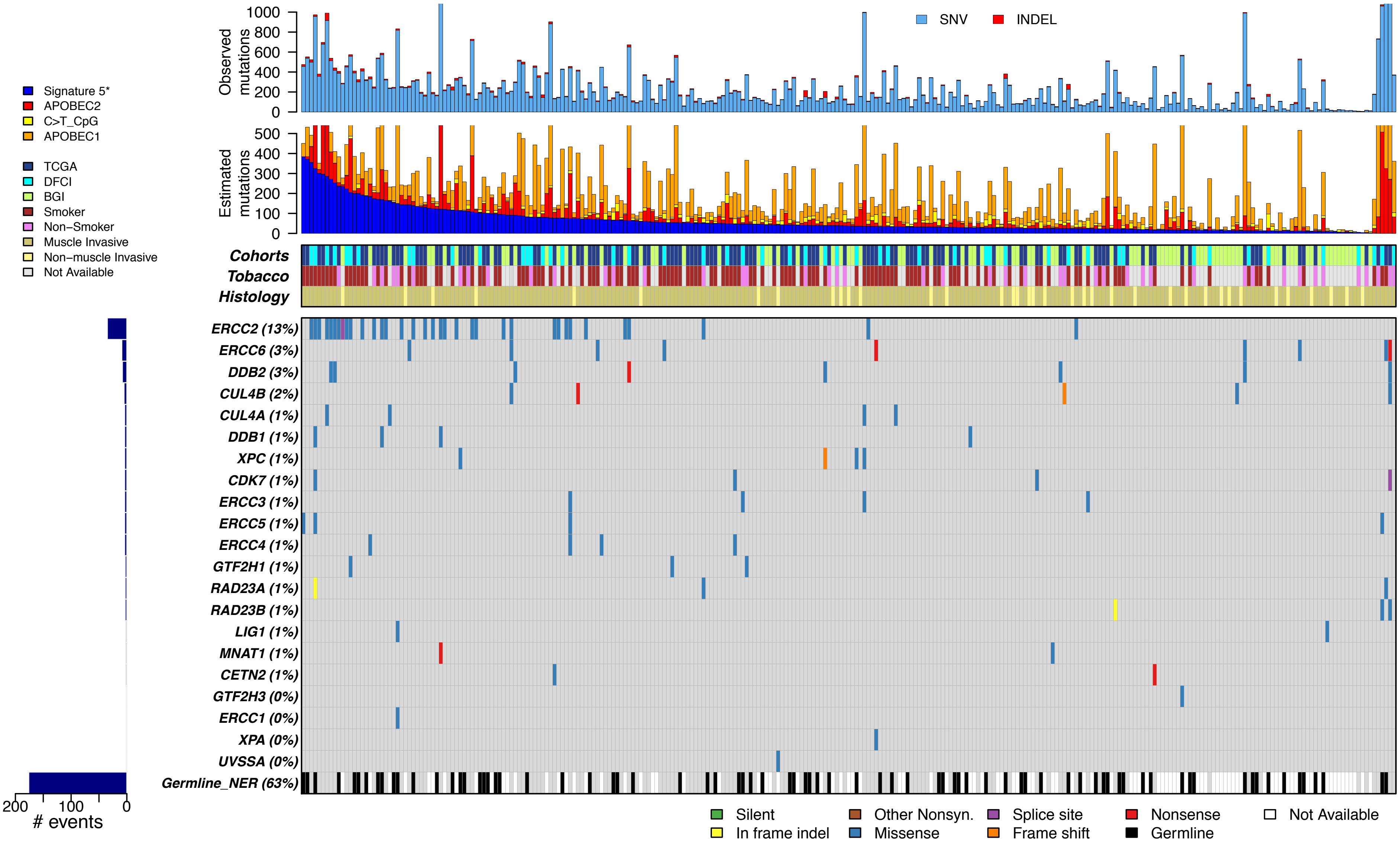
Supplementary Figure 8 Activities of other mutational signatures in mutant versus WT *ERCC2* tumors across cohorts. The median estimated number of mutations are shown in parentheses and one-sided p-values were computed using a permutation-based method that maintains the overall number of non-silent mutations per sample and per gene (**Methods**).[3] COMB-279: all cases across the three cohorts. COMB-MI-242: all muscle-invasive cases across the three cohorts. The C>T CpG signature is not shown for the BGI-99 cohort because this signature did not separate from signature 5* in the NMF analysis.

Supplementary Figure 9



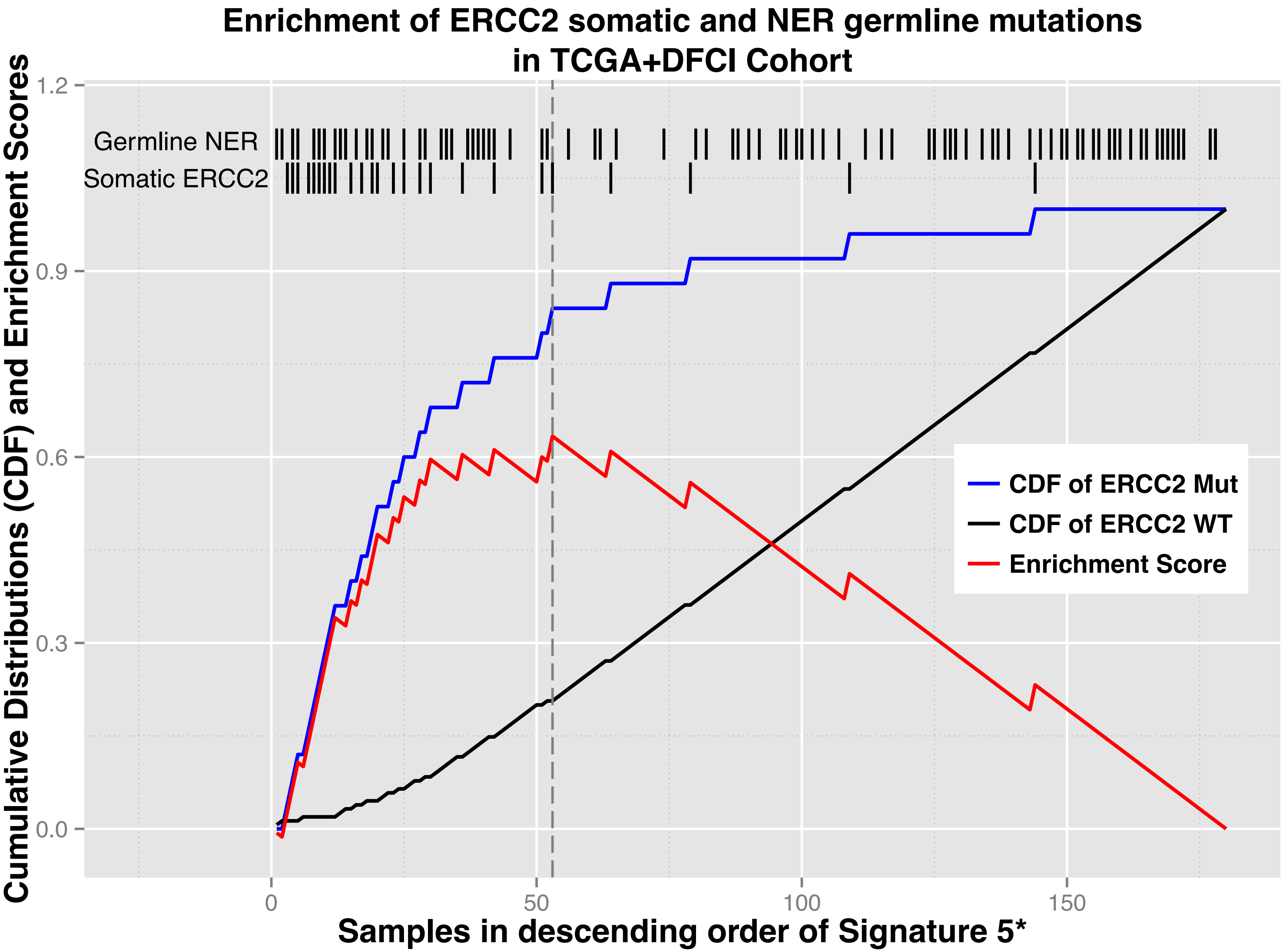
Supplementary Figure 9 Enrichment analysis for the APOBEC2 signature across genes. Although the p-value for *ERCC2* was <0.05 in each of the three cohorts, the Benjamini-Hochberg False Discovery Rate Q-value was <0.1 (shown in red) only in the combined cohorts (COMB-MI-242 and COMB-279).

Supplementary Figure 10

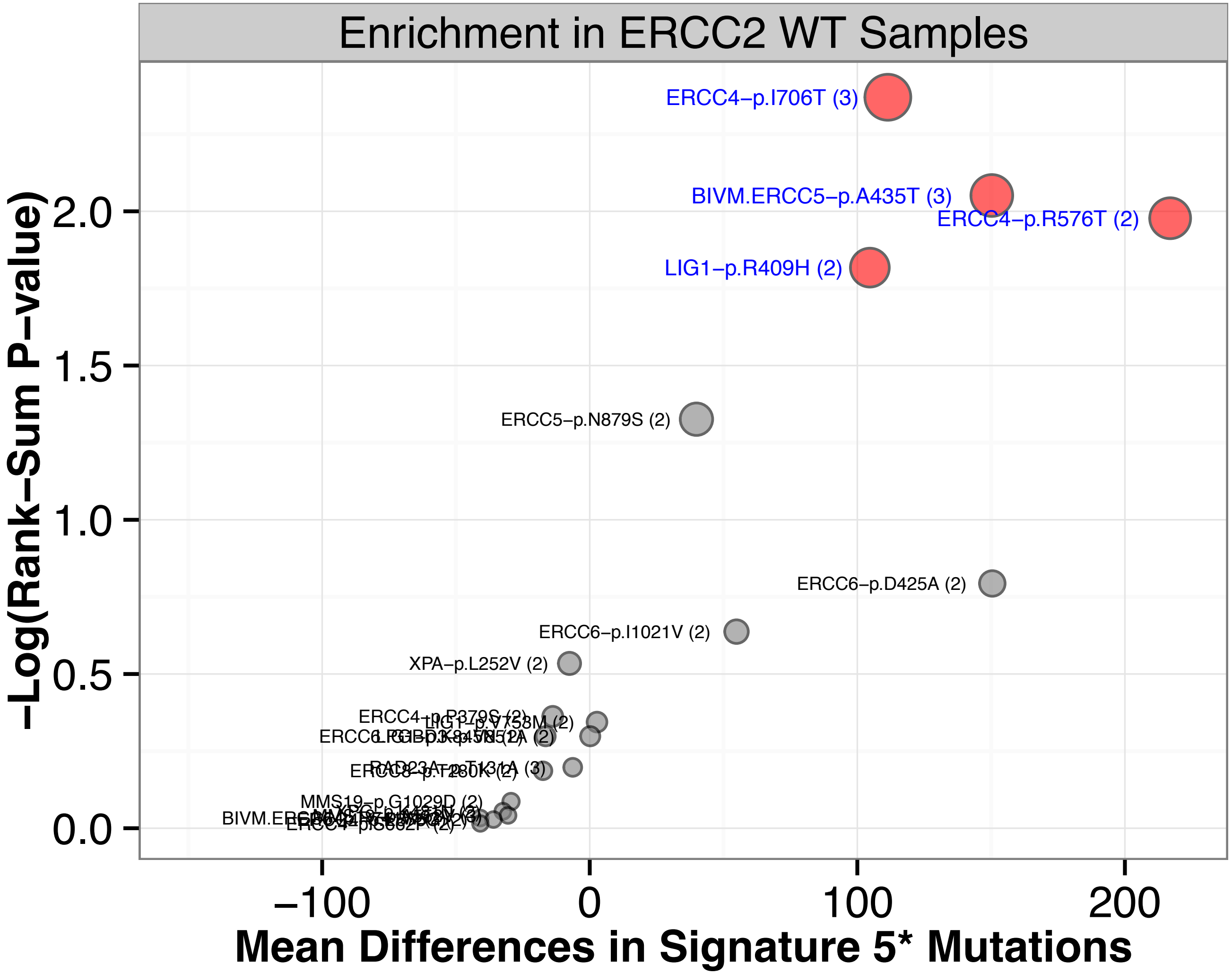


Supplementary Figure 10 Association of somatic and germline NER pathway events with signature 5* activity in the combined cohort (TCGA-130 + DFCI/MSK-50 +BGI-99). The figure is arranged similar to Figure 4 in the main text, but provides additional detail regarding events in NER pathway genes (see **Methods** for full list of NER pathway genes). Somatic mutations in non-*ERCC2* NER genes are rare, and there is no significant enrichment of mutations in any individual non-*ERCC2* NER gene or of the pathway as a whole (when *ERCC2* is excluded) among tumors with increased signature 5* activity. Rare germline NER variants (frequency <2% in the TCGA-130 + DFCI/MSK-50 cohort; see **Methods**) are displayed in a single track at the bottom of the figure.

Supplementary Figure 11a



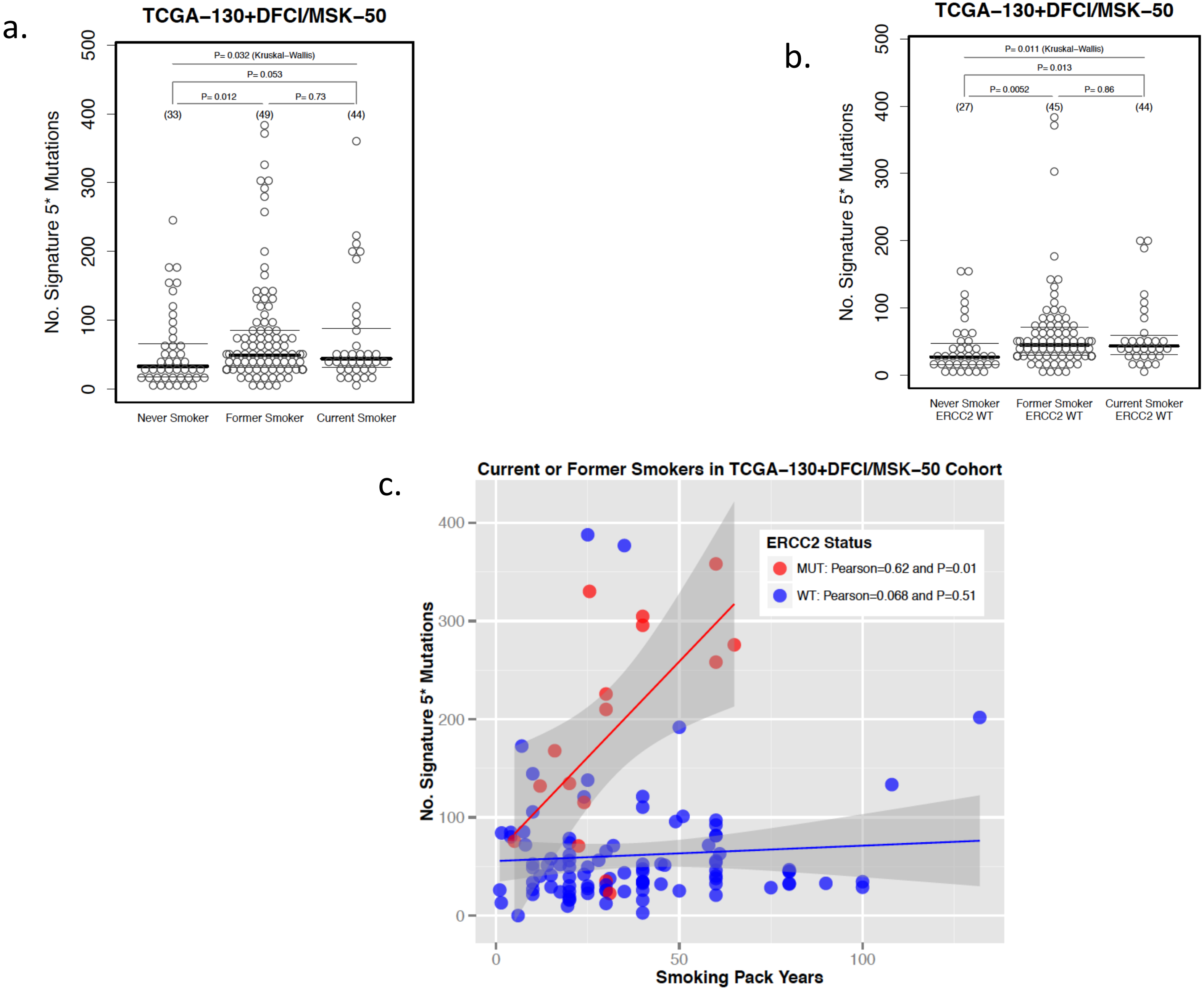
Supplementary Figure 11b



Supplementary Figure 11 Enrichment analyses for somatic and germline NER events. **(a)** Cumulative distributions and enrichment scores for somatic *ERCC2* mutations and germline NER pathway events. The dashed vertical line denotes the maximum enrichment score for somatic *ERCC2* mutations. Among the 32 WT *ERCC2* tumors with highest signature 5* activity (i.e., left of the dashed line), 19 (59%) had a germline NER variant, whereas among the 123 tumors with lower signature 5* activity (i.e., right of the dashed line), only 54 (44%) had a NER germline variant (p=0.086, Fisher's exact test). **(b)** Association of NER germline variants with signature 5* activity among WT *ERCC2* tumors from the TCGA-130 and DFCI-MSK-50 cohorts (germline data not available for BGI-99 cohort). Variant alleles present in >1 case but <2% of the population are shown (number of cases shown in parentheses), and alleles associated with significant enrichment (FDR<0.1) in signature 5* mutations are highlighted in red. Summary of the four significantly enriched alleles (annotations taken from ExAC [5]):

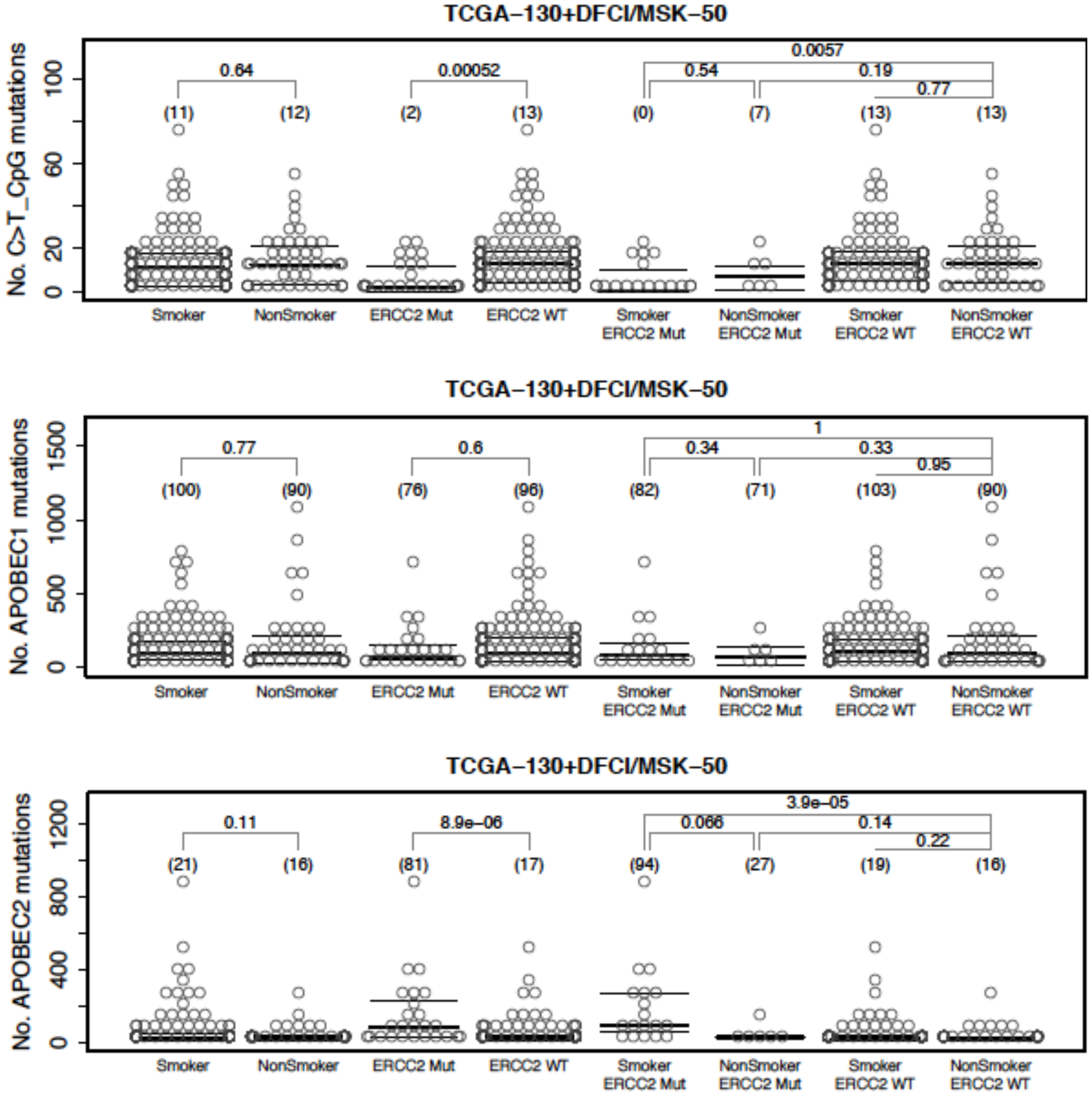
Variant	Polyphen2	SIFT	Population frequency
ERCC4-p.I706T	probably damaging	deleterious	0.0014
ERCC4-p.R576T	benign	deleterious	0.00054
LIG1-p.R409H	possibly damaging	deleterious	0.014
BIVM-ERCC5-p.A435T	no annotation		

Supplementary Figure 12



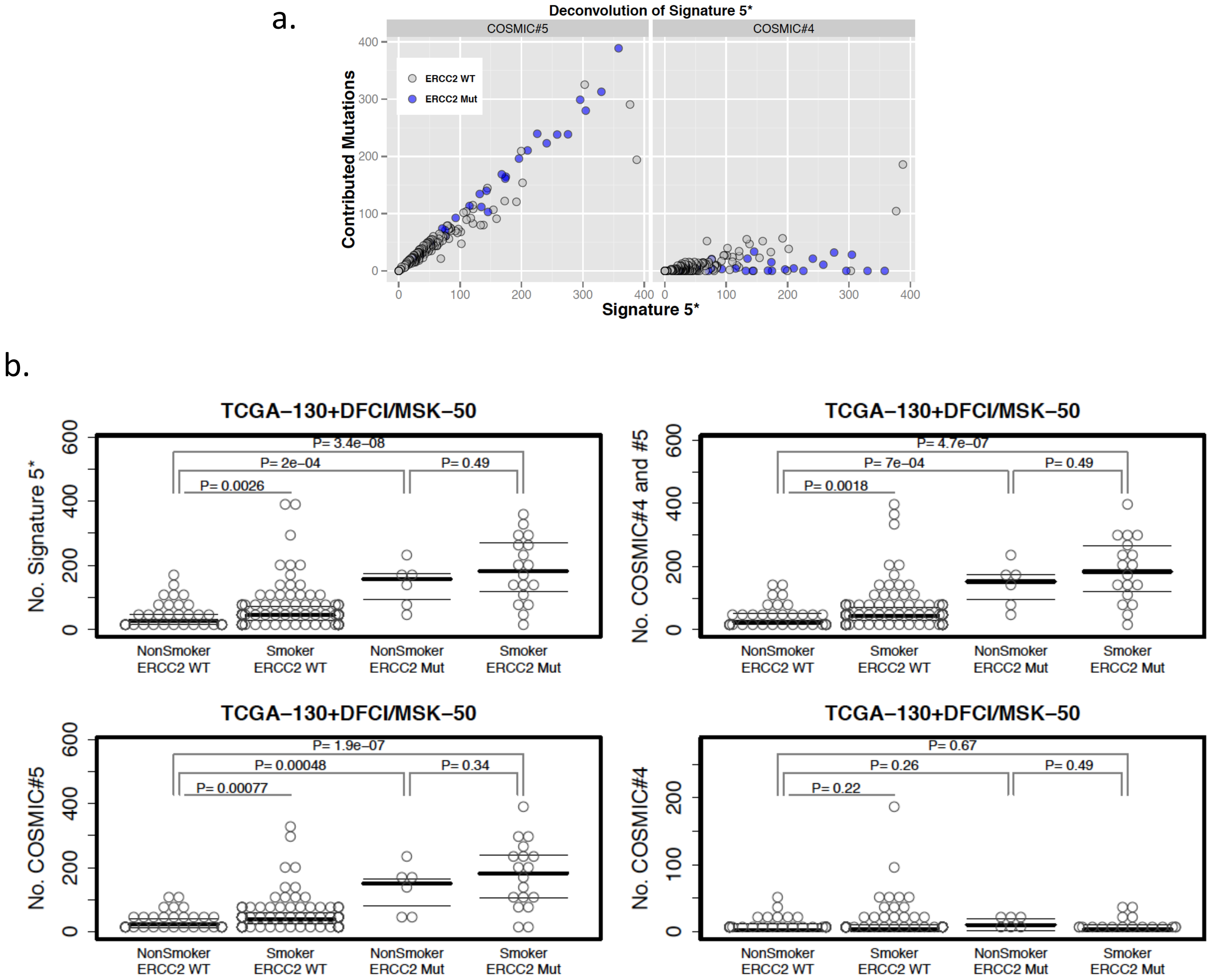
Supplementary Figure 12 Effect of current smoking status and smoking intensity on signature 5* activity in the combined TCGA-130 + DFCI/MSK-50 cohort. **(a)** There was no difference in the estimated number of signature 5* mutations in current versus former smokers. The estimated number of signature 5* mutations is shown in parentheses. P-values were calculated using the Wilcoxon rank-sum test. **(b)** There was also no difference in estimated number of signature 5* mutations in current versus former smokers when only WT *ERCC2* cases were considered. **(c)** There was a correlation between smoking intensity (measured in pack-years exposure) and signature 5* activity for *ERCC2* mutated cases but not WT *ERCC2* cases.

Supplementary Figure 13



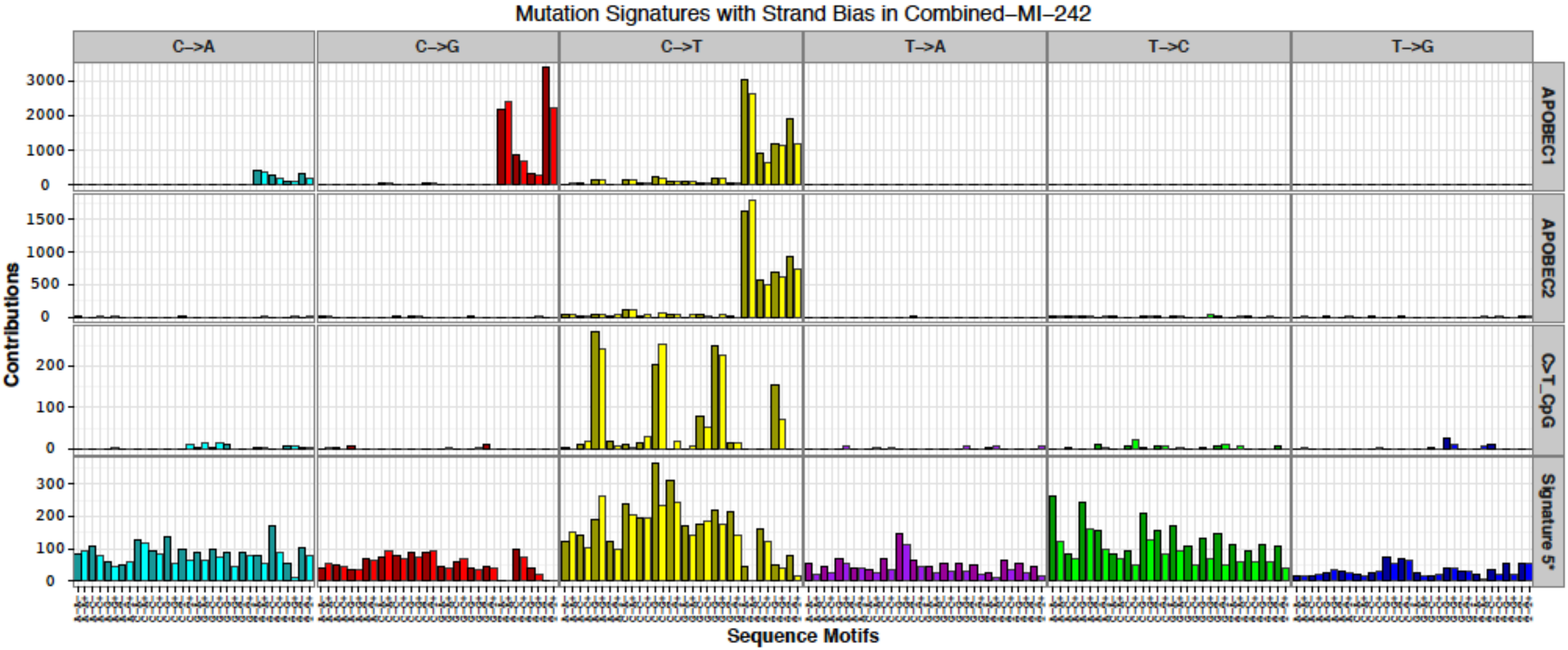
Supplementary Figure 13 Association of smoking with other mutational signatures in the combined TCGA-130 + DFCI/MSK-50 cohort. There was no association between smoking status and activity of any of the other mutational signatures identified in the cohorts.

Supplementary Figure 14

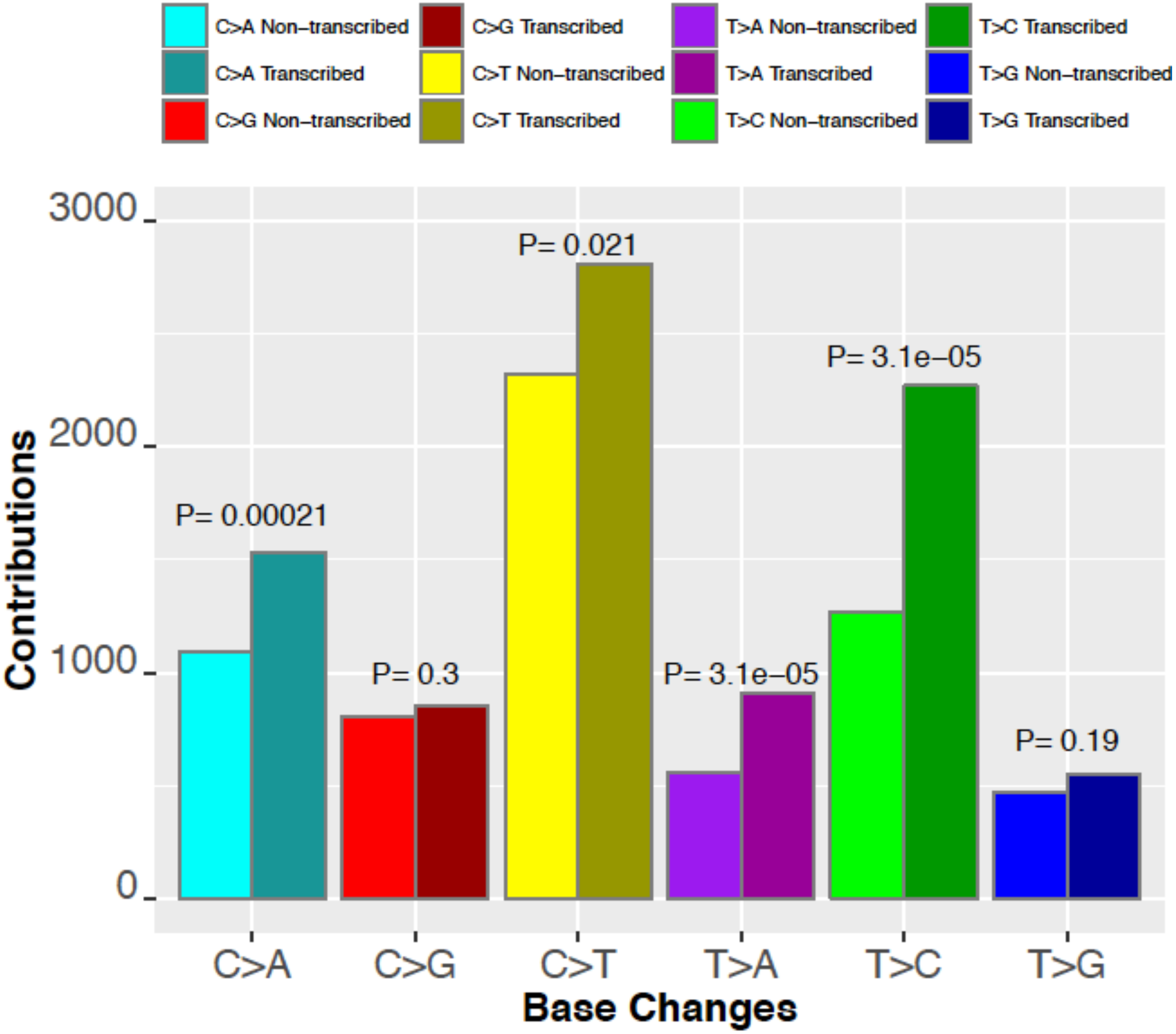


Supplementary Figure 14 Relationship between signature 5* and COSMIC signatures 4 and 5. The contributions of COSMIC signatures 4 and 5 to signature 5* were determined by deconvoluting signature 5* mutations into COSMIC signature 4 and COSMIC signature 5 components (**Methods**). **(a)** Nearly all signature 5* activity is attributable to activity of COSMIC signature 5, with only a small contribution from COSMIC signature 4. The two samples with strongest contribution from COSMIC signature 4 were WT *ERCC2* cases. **(b)** The difference in signature 5* activity between smokers and non-smokers is due to differences in activity of COSMIC signature 5 rather than COSMIC signature 4.

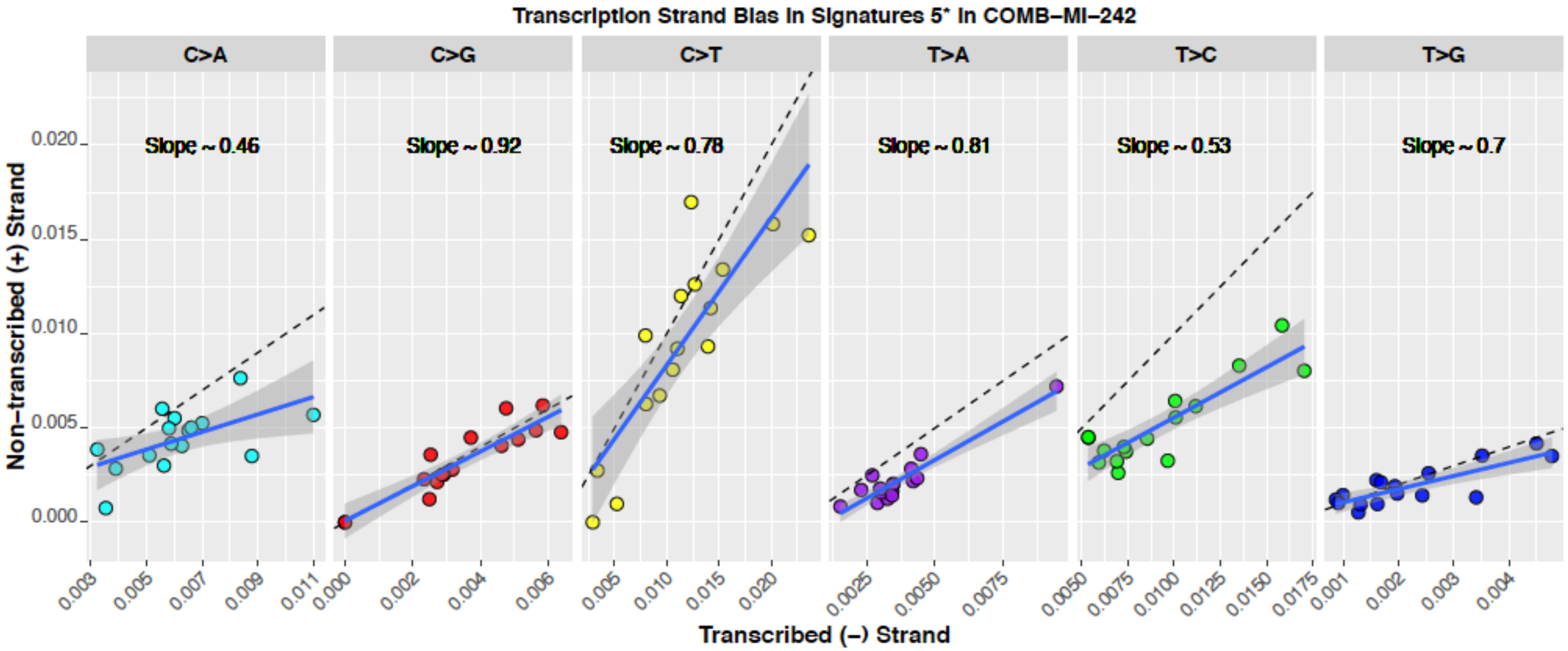
Supplementary Figure 15a



Supplementary Figure 15b

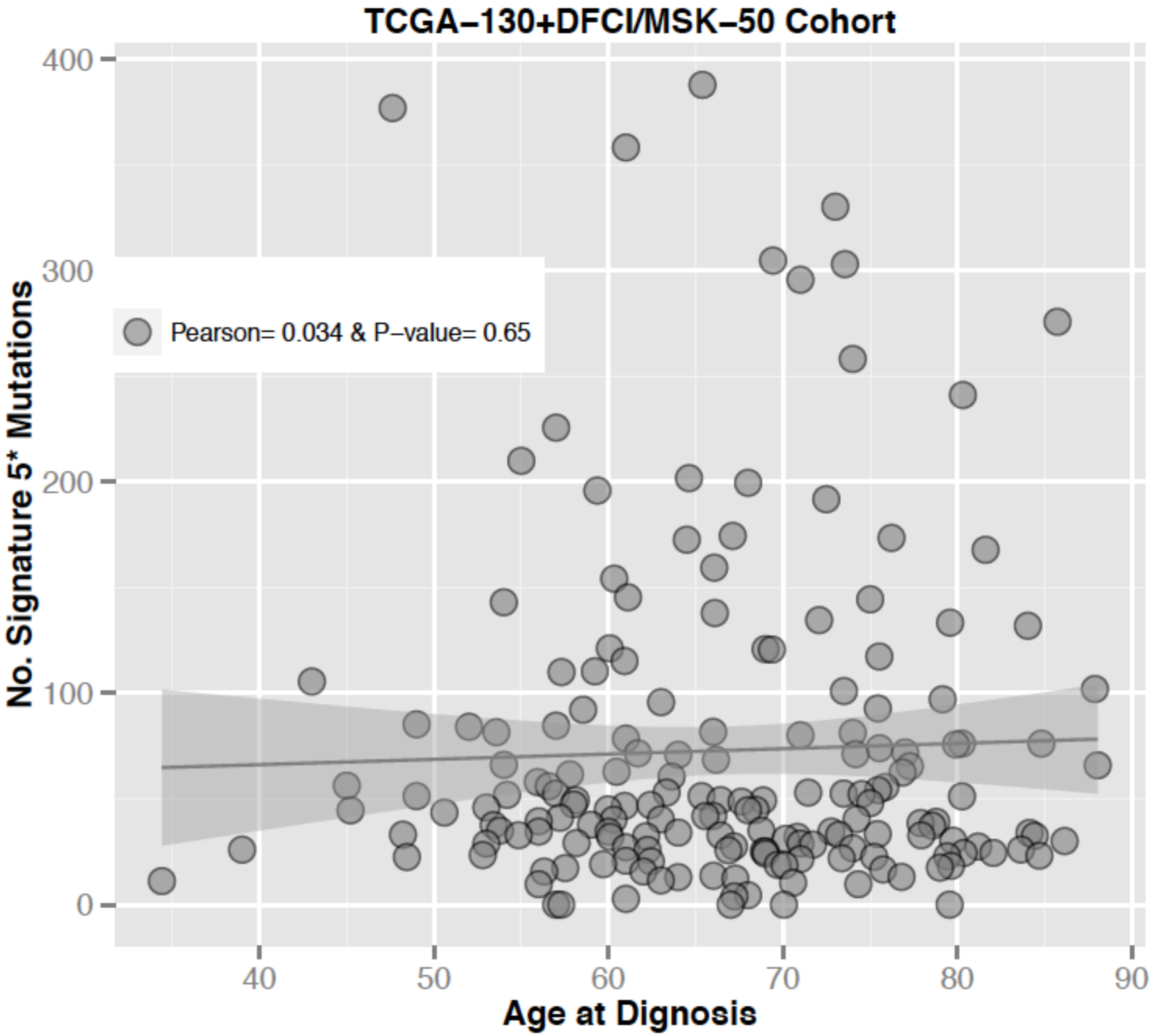


Supplementary Figure 15c



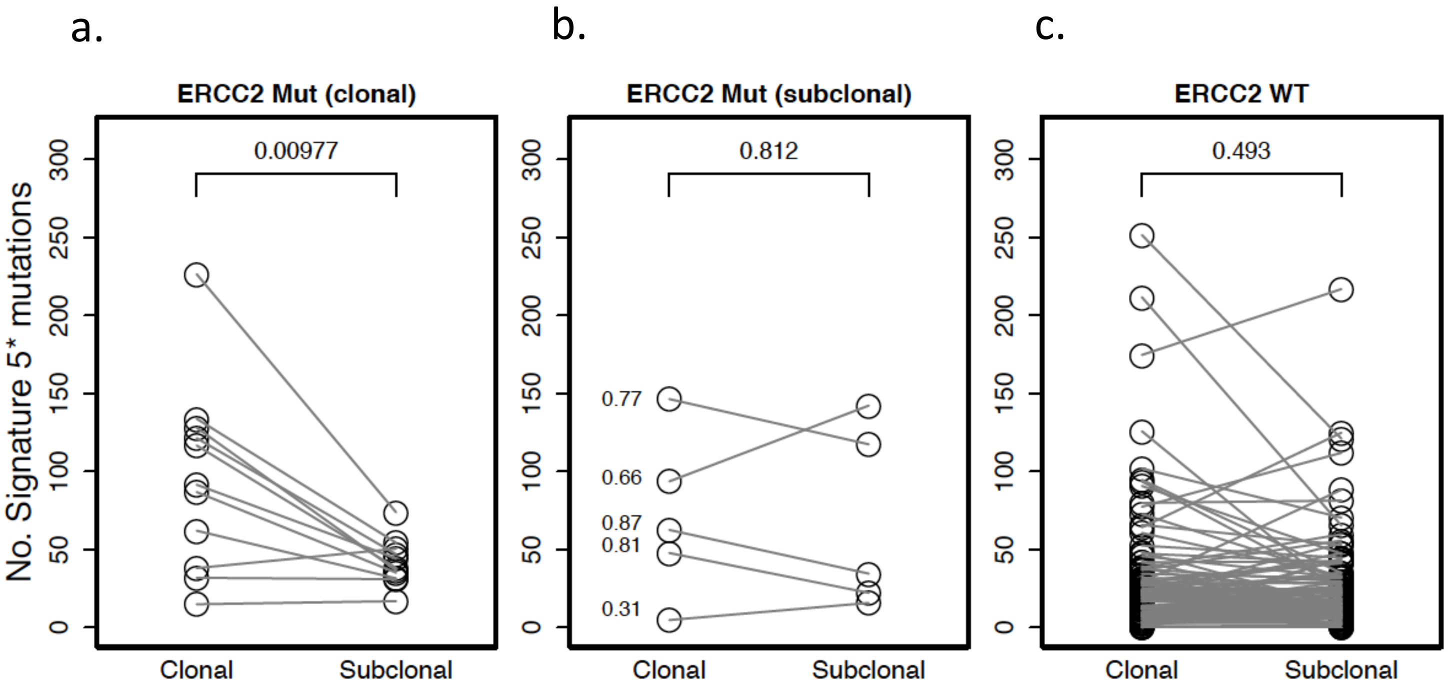
Supplementary Figure 15 Strand asymmetry analysis. For all muscle-invasive tumors across the three cohorts (COMB-MI-242), the Bayesian NMF analysis was repeated while considering mutations on the transcribed and non-transcribed strands separately (i.e., 192 rather than 96 trinucleotide mutational contexts). **(a)** Estimated number of mutations on transcribed (shaded colors) and non-transcribed strands (non-shaded colors) for each of the four signatures. **(b)** Summary of single base changes on the transcribed versus non-transcribed strand. P-values were computed using the pair-wise Wilcoxon rank-sum test. **(c)** The activity on the transcribed versus non-transcribed strand is displayed for each of the six possible base pair changes. Signature 5* exhibits a transcriptional strand bias that is strongest for T>C and C>A changes.

Supplementary Figure 16



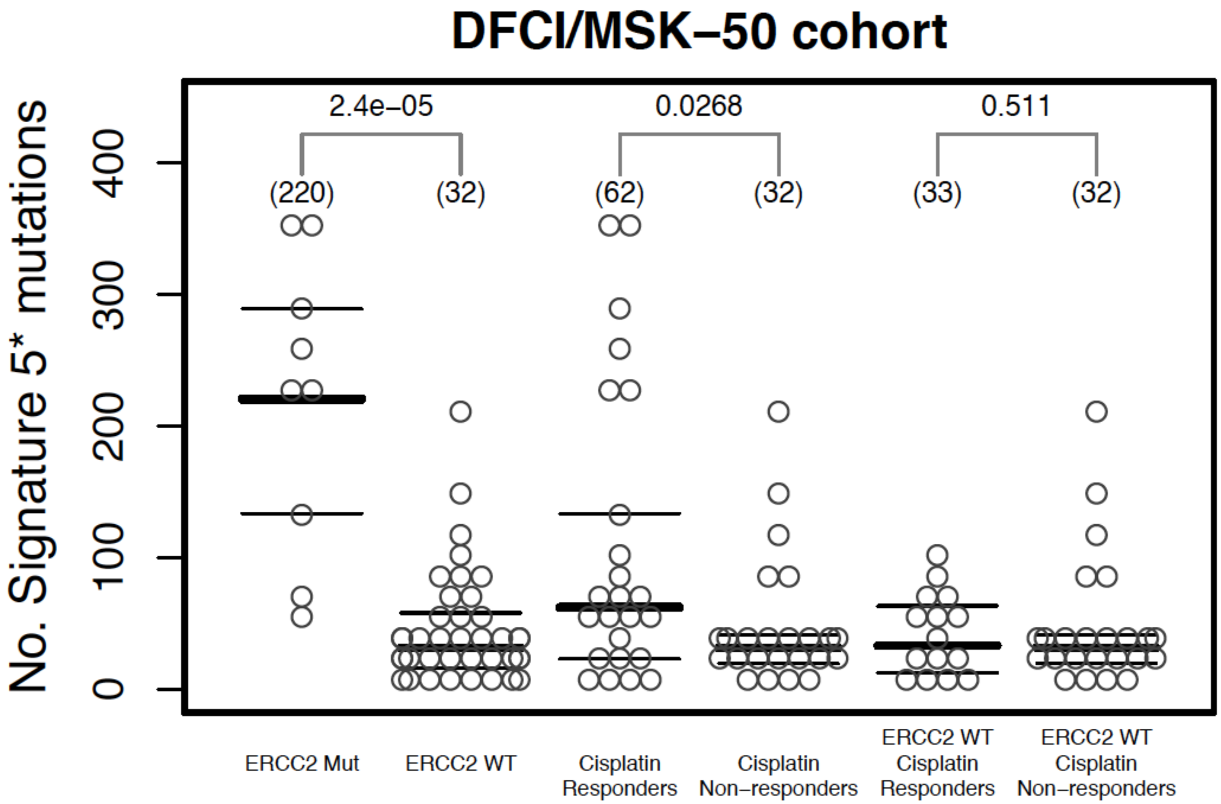
Supplementary Figure 16 Association between patient age and signature 5* activity in tumors from the TCGA-130 and DFCI/MSK-50 cohorts (the two cohorts with available age data). As has been previously described for COSMIC signature 5 in urothelial cancer, there was no association between age at diagnosis and signature 5* activity in the urothelial tumors analyzed here (P=0.65).[6]

Supplementary Figure 17



Supplementary Figure 17 Clonality of signature 5* mutations in WT versus mutant *ERCC2* tumors. For each tumor with an *ERCC2* mutation, the number of clonal (defined as probability [cancer cell fraction$\geq$0.95]$>$0.5) and subclonal signature 5* mutations are shown. Gray lines connect counts from the same tumor. **(a)** Tumors with a clonal *ERCC2* mutation have significantly more clonal than subclonal signature 5* mutations. **(b)** There is no significant difference in the number of clonal versus subclonal signature 5* mutations in tumors with a subclonal *ERCC2* mutation. **(c)** There is also no significant difference in the number of clonal versus subclonal signature 5* mutations in tumors with WT *ERCC2*. P-values were calculated using the pairwise Mann-Whitney test.

Supplementary Figure 18



Supplementary Figure 18 Association between signature 5* activity and cisplatin response in the DFCI/MSK-50 cohort (the only cohort with cisplatin response data available). Median estimated number of signature 5* mutations are shown in parentheses and p-values were calculated using the Wilcoxon rank-sum test. There were a significantly higher number of signature 5* mutations in cisplatin responders versus non-responders; however, this difference was not significant when only WT *ERCC2* tumors were considered.

Supplementary Tables (see separate files)

Supplementary Table 1 Summary of Urothelial Cancer Cohorts.

Supplementary Table 2 Numerical Representation of Signature 5* Across Cohorts.

Supplementary Table 3 Summary of Mutational Signature Contributions, *ERCC2* Mutational Status, and Smoking Status for All Cases.

Supplementary Table 4 Comparison of Mutational Signatures in Urothelial Tumor Cohorts to COSMIC Mutational Signatures.

Supplementary Table 5 Comparison of Signature 5* Among Urothelial Tumor Cohorts.

Supplementary References

1. COSMIC: Catalogue of Somatic Mutations in Cancer. [cited 2015 October 25]; Available from: http://cancer.sanger.ac.uk/cosmic/signatures.
2. Alexandrov, LB, Nik-Zainal, S, Wedge, DC, Aparicio, SA*, et al.* Signatures of mutational processes in human cancer*. Nature* 2013;**500**:415-21.
3. Strona, G, Nappo, D, Boccacci, F, Fattorini, S, San-Miguel-Ayanz, J. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals*. Nat Commun* 2014;**5**:4114.
4. Kamburov, A, Lawrence, MS, Polak, P, Leshchiner, I*, et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures*. Proc Natl Acad Sci U S A* 2015;**112**:E5486-95.
5. Consortium, EA. Analysis of protein-coding genetic variation in 60,706 humans*. BioRx* 2015;
6. Alexandrov, LB, Jones, PH, Wedge, DC, Sale, JE*, et al.* Clock-like mutational processes in human somatic cells*. Nat Genet* 2015;